

On Support Relations and Semantic Scene Graphs

Wentong Liao¹, Michael Ying Yang², Hanno Ackermann¹ and Bodo Rosenhahn¹

Abstract—Rapid development of robots and autonomous vehicles requires semantic information about the surrounding scene to decide upon the correct action or to be able to complete particular tasks. Scene understanding provides the necessary semantic interpretation by semantic scene graphs. For this task, so-called support relationships which describe the contextual relations between parts of the scene such as floor, wall, table, etc, need be known. This paper presents a novel approach to infer such relations and then to construct the scene graph. Support relations are estimated by considering important, previously ignored information: the physical stability and the prior support knowledge between object classes. In contrast to previous methods for extracting support relations, the proposed approach generates more accurate results, and does not require a pixel-wise semantic labeling of the scene. The semantic scene graph which describes all the contextual relations within the scene is constructed using this information. To evaluate the accuracy of these graphs, multiple different measures are formulated. The proposed algorithms are evaluated using the NYUv2 database. The results demonstrate that the inferred support relations are more precise than state-of-the-art. The scene graphs are compared against ground truth graphs.

I. INTRODUCTION

Scene understanding is a popular but challenging topic in computer vision, robots and artificial intelligence. It can be roughly divided into object recognition [1], layout estimation [2], and physical relations inference [3]. Traditional scene understanding mainly focuses on object recognition and has achieved great developments, especially by recent developments in deep learning [4]. Exploring more vision cues like contextual and physical relations between objects is becoming the topic of great interest in the computer vision community. In many robotic applications as well, knowledge about relations between objects are necessary for a robot to finish its task. For example, for a robot to take a newspaper from under a cup, it must first lift the cup, and then put it back or place it somewhere else.

A semantic scene graph is an effective tool for representing physical and contextual relations between objects and scenes. In [5] it was proposed to use a semantic scene graph in a robotic application. Scene graphs have also been used in different applications [6], [7] and [8]. However, in most existing works, a scene graph is regarded as input for scene understanding.

Our goal in this work is to infer reasonable support relations and then to generate a semantic scene graph. To accurately estimate support relations, we propose a framework

based on object detection and contextual semantics instead of pixelwise segmentation or 3D cuboids which are used in previous methods for inferring support relations.

With the information achieved from scene recognition, object recognition, attribute recognition, support estimation and relative spacial estimation, a semantic graph is inferred to describe the given scene. Additionally, we introduce some metrics to evaluate the quality of a generated semantic graph, an issue so far not considered. An overview of our approach is illustrated in Fig. 1.

We analyze our method on the benchmark dataset NYUv2 of cluttered room scenes. The results show that our algorithm outperforms the state-of-the-art for support inference. Quantitative and qualitative comparisons with ground truth scene graphs show that the estimated graphs are accurate.

To summarize, our contributions are:

- We propose a new method for inferring more accurate support relations compared to previous works [9], [10].
- Neither pixel-wise semantic labelings nor 3D cuboids are necessary.
- We introduce a way how to construct semantic scene graphs and assess the quality
- Ground truth scene graphs of the NYUv2 dataset are provided to the scientific community
- A convenient GUI tool for generating ground truth graphs will be made available¹

This paper is structured as follows: related work is discussed in Sec. II. Object recognition and segmentation, features for scene and support relations classification are shortly explained in Sec. III. In Sec. IV, the model for support inference is proposed. How a scene graph is inferred can be explained in Sec. V. Experimental results of the proposed framework are shown in Sec. VI. Finally, a conclusion in Sec. VII summarizes this paper.

II. RELATED WORK

Justin et al. [7] proposed to use scene graphs as queries to retrieve semantically related images. Their scene graphs are manually generated by the Amazon Mechanical Turk, which literally is expensive. Prabhu and Venkatesh [11] constructed scene graphs to represent the semantic characteristics of an image, and used it for image ranking by graph matching. Their approach works on high-quality images with few objects. Lin et al. [12] proposed to use scene graphs for video search. Their semantic graphs are generated from text queries using manually-defined rules to transform parse trees,

¹Wentong Liao, Hanno Ackermann and Bodo Rosenhahn are with Institute of Information Processing Leibniz University Hannover, Germany

²Michael Ying Yang (Corresponding Author) is with University of Twente-ITC, Netherlands (michael.yang@utwente.nl)

¹More information on the tool can be found in the supplementary material of this paper. This information will be provided on the authors' homepage.

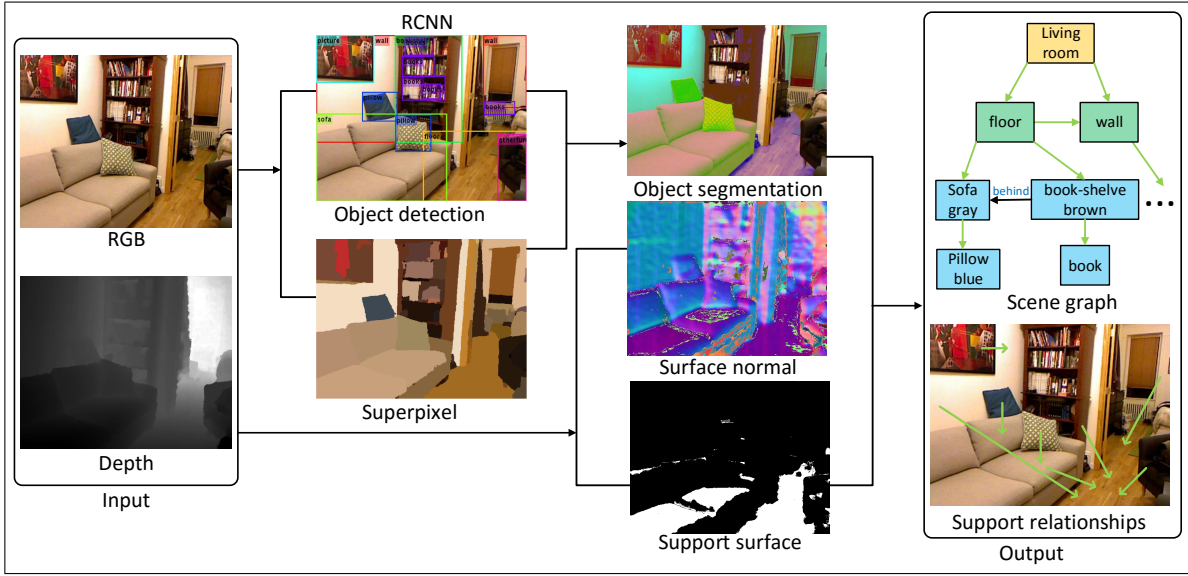


Fig. 1. **Overview.** Given an RGBD image, we use a CNN for object detection. A superpixel map is also computed. *Bounding boxes* of detected objects and the *superpixel map* are used to segment objects. Parallel to this process, surface normals, support surface and 3D point cloud aligned to the room are computed. Physical support between objects are estimated and a scene graph is inferred. Green arrows indicate the relation from the supported object to the surface that supports it.

similar as [8]. Using a grammar, Liu et al. [13] proposed to learn scene graphs from ground truth graphs of synthetic data. Then they parsed through a pre-defined segmentation of a synthetic scene so as to create a graph that matches the learned structure. None of these works objectively assess the quality of scene graph hypotheses compared with ground truth graphs. However, reasonable measures for this problem are important especially after the publication of the *Visual Genome* dataset [14].

Physical relations between objects to help image or scene understanding have been investigated in [9], [10], [15], [16], and [17]. Pixel-wise segmentation and 3D volumetric estimation are two major methods for this task. [9], [10] used pixel-wise segmentations to analyze support relations in challenging cluttered indoor scenes. They both ignored the contextual knowledge provided by the scene. Silberman et al. [9] ignored small objects and the physical constraints while Xue et al. [10] set up simple physical constraints. A typical examples of 3D cuboid based method is [15]. Jia et al. estimated the 3D cuboids to capture spatial information of each object using RGBD data and then reason about their stability. However, stability and support relations are inferred in tiny images with few objects.

For the part of support relations inference in this paper is mostly related to [9], [10]. However, we integrate physical constraints and prior support knowledge between object classes into our approach for extracting more accurate support relations. Furthermore, we do not operate pixelwise segmentation for object extraction. Finally, our framework generates a semantic graph to interpret the given image. Objective measures for accessing the quality of constructed graphs are proposed.

III. OBJECT DETECTION AND CLASSIFICATION

Recently, deep learning based algorithms have shown great success in object detection and classification tasks [18], [4], and [19]. Here, the RCNN framework proposed by Gupta et al. [19] is applied to recognize objects in an image, which utilizes HHA, abbr. of horizontal displacement (depth), height and angle of the pixels local surface normal, representations to enhance object classification ability. In the RCNN, a pool of candidates is created and then each is indicated by a bounding box and a confidence score sb . Then, a class specific CNN assigns each proposal a classification score sc . Finally, a non-maximum suppression is used to remove overlapping detections and obtain the final object proposals. Different from the original work, a weighted score sw_i is used in the final step to obtain better bounding boxes of detected objects

$$sw_i = sb_i + w * sc_i, \quad i \in 1 \dots N, \quad (1)$$

where w is the weight factor and N is the total number of detected objects. Fig. 2(a) shows an example of proposal decision. The proposals (in green and red bounding boxes respectively) are two out of several proposals that are classified as table and have large intersection. The red bounding box is decided using the weighted score in the non-maximum suppression step while the green box is the result of [19]. It can be seen that the red one covers the table more accurately than the green one. This result is important in the following step of estimating relative positions between objects.

A. Object Segmentation

Correct separation of foreground objects from the background in the bounding box is critical for estimating accurate support relationships. Many approaches can effectively

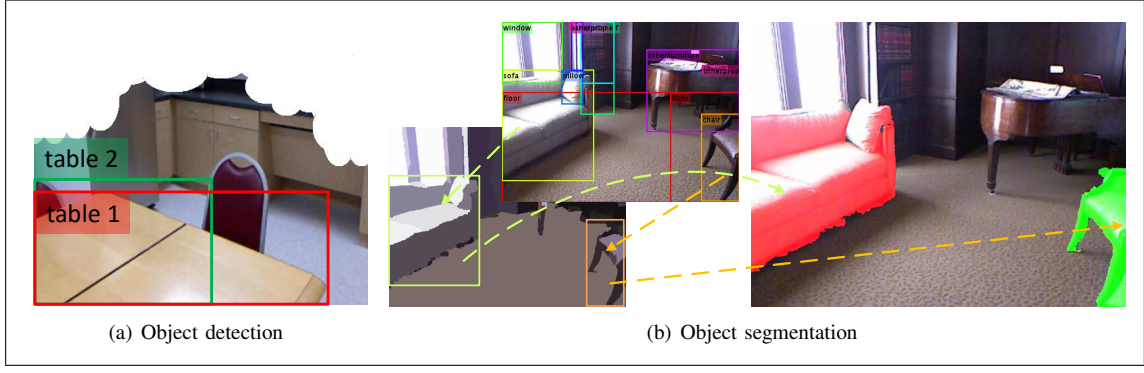


Fig. 2. Examples of object detection (a) and object segmentation from bounding box and superpixel map (b). (a) two of the proposals with large intersection are classified as table. (b) The left two images show detected objects in bounding boxes from in RGB image (left-up) and corresponding superpixel map (left-down). The right image illustrates the results of object segmentation (only shows sofa and chair).

complete this task, such as Grabcut [20], [21], semantic segmentation [22] and instance segmentation [19]. But they are very computation time costly: Grabcut takes minutes for each bounding box, and deep learning-based approaches need several days for training and still several minutes for each input image. Furthermore, they require high performance GPUs which is a limitation for practical applications, e.g. robots. We propose a simple and fast method to segment objects based on superpixel maps.

Starting with the smallest (area) bounding box and continuing in ascending order, superpixels are assigned to bounding boxes if more than a certain ratio (in this paper: 80%) of their area is within the box. Fig. 2(b) shows an example of segmentation result of our method: the sofa and chair are well separated from the background. This method is very efficient in terms of computation time and power.

The superpixel maps used in this step are produced by the RCNN during the object recognition process. In other words, no additional computation is required to generate superpixel map for each image.

B. Scene and support relations classification

Indoor scene category is an important auxiliary information for object recognition, for instance, a bathtub is impossible in a living room, as shown in Fig. 3. In this paper, we use the spatial pyramid pooling of [23] as feature for this task and use logistic regression classifier to make probabilistic prediction.

We furthermore benefit from using this method, which the SIFT descriptors generated in this process are used as the feature for classifying support relations, then computation is saved by extracting specific features. A logistic regression classifier D_{SP} is trained with features $F_{i,j}^{SP}$ associating with support label $y^s \in \{1, 2, 3\}$ to indicate object j supports object i from below, behind, and no support relationship, respectively. Note that the features are asymmetric, i.e. $F_{i,j}^{SP}$ is not for judging if object i supports object j .

C. Coordinate Alignment

A suitable coordinate system is necessary for correctly estimating object positions. Therefore, the image coordinates

need to be aligned with the 3D room coordinates first. We find the principle directions ($\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$) of the aligned coordinate based on the Manhattan world assumption [24]: most of the visible surfaces are located along one of three orthogonal directions. The "wall" or "floor/ground" (We note the floor as ground for convenient interpretation in the follows of this paper) detected by RCNN (if yes) indicates the horizontal or vertical direction of the scene. This useful cues are embodied to our method for coordinate alignment. Each pixel has image coordinates (u, v) , 3D coordinates (X, Y, Z) , and the local surface normal (N_x, N_y, N_z) . As discussed in [9], straight lines are extracted from images and the mean-shift modes of surface normals are computed. For each line that is very closed to Y direction, two other orthogonal candidates are sampled for computing the score as follows:

$$S(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = \sum_{j=1}^3 (SN_j + SL_j) \quad j = 1, 2, 3 \quad (2)$$

$$SN_j = \frac{w_N}{N_N} \sum_i^{Num_N} \exp\left(-\frac{(N_i * V_j)^2}{\sigma^2}\right) + \mathbb{I}(y_{N_i})P_{N_i} \quad (3)$$

$$SL_j = \frac{w_L}{N_L} \sum_i^{Num_L} \exp\left(-\frac{(L_i * V_j)^2}{\sigma^2}\right). \quad (4)$$

Here, $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ indicate the three principal directions, N_i the normal of a pixel, L_i the direction of a straight line, Num_N and Num_L the number of points and lines on each surface, respectively, and w_N and w_L the weights of the 3D normals and line scores, respectively. $\mathbb{I}(y_{N_i}) = 1$, if the region which includes the pixel N_i is "ground" or "wall" and P_{N_i} is the corresponding predicted probability, else $\mathbb{I}(y_{N_i}) = 0$. This term favors the candidate that is most perpendicular to the ground or wall surface to be chosen as one of the principle direction and further ensures the ground to be the lowest surface. The candidates ($\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$) which have the maximal score are chosen as the aligned coordinate system and the one of the three directions which is closest to the original Y direction is chosen as \mathbf{v}_y . Then the image coordinate is aligned to $(\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z)$.

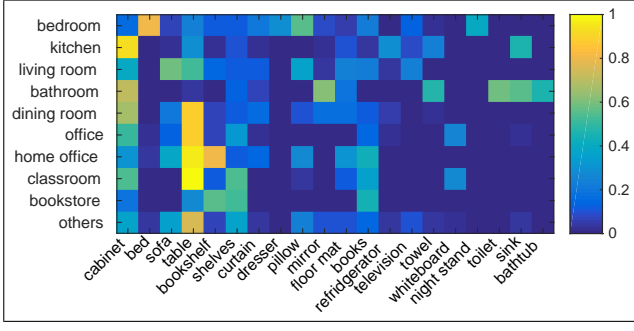


Fig. 3. Prior knowledge of specific class object presenting in specific scene type. Not all of the object categories are shown here.

IV. MODELING SUPPORT RELATIONSHIPS

Given an image with N detected objects with class labels $\mathbf{C} = \{C_1, \dots, C_N\}$, then (a) the visible supporter of object C_i is denoted by $S_i \in \{1 \dots N\}$, (b) $S_i = N + 1$ indicates that object C_i is supported by an invisible object and (c) $S_i = \text{ground}$ means that C_i is the ground and does not need support. The supporting type of C_i is encoded as $ST_i = 1$ for being supported from behind and $ST_i = 0$ for the support from below.

The following assumptions are used in our model: Every object is either (a) supported by another detected object next to it in the image plane, in which case $S_i \in \{1 \dots N\}$, (b) supported by an object not detected or invisible in the image plane, or $S_i = N + 1$, (c) it is ground itself which requires no support $S_i = \text{ground}$.

In practice, the object classes and physical factors (e.g. physical rationality and common sense) constrain the support relations of indoor scenes. For example, the window behind the sofa in Fig. 2(b) is supported by the wall rather than other objects or parts of the scene. However, without such common sense rules, it is more likely to be supported by the sofa. To infer more accurately support relationships, the object classes information and some constraints are added to our model.

We infer the support relationships similarly as in [9]: the most probable joint assignment of support object $\mathbf{S} = \{S_1 \dots S_N\}$, support type $ST \in \{0, 1\}$ and object classes $\mathbf{C} = \{C_1, \dots, C_N\}$

$$\{\mathbf{S}^*, \mathbf{ST}^*, \mathbf{C}^*\} = \arg \min_{\mathbf{S}, \mathbf{T}, \mathbf{C}} E(\mathbf{S}, \mathbf{T}, \mathbf{C}). \quad (5)$$

The energy of our model in Eq. (5) is divided into four parts: the support energy E_{SP} , the object classification E_C , and the physical constraint energy E_{PC} . The total energy function is formally defined as:

$$E(S, ST, SC) = E_{SP}(S, ST) + E_C(C) + E_{PC}(S, ST, C), \quad (6)$$

where

$$E_{SP}(S, ST) = - \sum_i^N \log(D_{SP}(F_{i,j}^{SP} | S_i, ST_i)), \quad (7)$$

$$E_C(C) = - \sum_i^N \log\{P_{C_i} P(C_i | SC) D_{SC}(F^{SC} | SC)\}. \quad (8)$$

Here, D_{SP} is the trained support classifier, $F_{i,j}^{SP}$ are the support features for C_j supporting C_i , P_{C_i} is the object category probability of C_i predicted by the RCNN, $P(C_i | SC)$ is the probability of object class C_i being present in the scene. D_{SC} is the trained scene classifier and $F_{i,j}^{SC}$ are the features. Fig. 3 shows the prior knowledge of 20 object classes in the dataset. For instance, $P(\text{bed} | \text{bedroom}) = 0.9$ means that 90% of images taken in bedroom do have a bed.

The physical constraint energy E_{PC} consists of several items: (1) The **Object class constraint** C_C : This object class constraint is imposed onto the support relations of a given object. For any support object, its lowest point should not be higher than the highest points of the supported object. The ground needs no support and must be the lowest points in the aligned coordinate system,

$$C_C(S_i, C_i) = \begin{cases} -\log P_{C_{S_i}, C_i}^{SP}, & \text{if } C_i \neq \text{ground AND } H_{S_i}^b \leq H_i^t, \\ -\log P(C_i), & \text{if } C_i = \text{ground AND} \\ & H_j^b > H_i^b, \forall C_j \\ \infty, & \text{otherwise.} \end{cases} \quad (9)$$

Here, H_i^b and H_i^t are the lowest and highest points in aligned 3D coordinates of object C_i , respectively, and $P_{C_{S_i}, C_i}^{SP}$ encodes the prior of object class C_{S_i} supporting object class C_i (but not vice versa).

(2) The **Distance constraint** C_{dist} : for any object, its supporter must be adjacent to it to satisfy the principle of physical stability. Formally, the distance constraint is defined as:

$$C_{dist}(S_i, C_i, ST_i) = \begin{cases} (H_i^b - H_{S_i}^t)^2, & \text{if } ST_i = 1, \\ V(S_i, C_i), & \text{if } ST_i = 0, \end{cases} \quad (10)$$

where $V(S_i, C_i)$ is the minimum horizontal distance from C_i to its supporter S_i .

(3) The **Support constraint** C_{SPC} : Besides the ground, all detected objects must be supported, and no object is lower than the ground. This constraint is formally defined as:

$$C_{SPC}(S_i, C_i) = \begin{cases} \infty, & \text{if } S_i = \text{ground AND } C_i \neq \text{ground} \\ \infty, & \text{if } C_i = \text{ground AND } H_j^b < H_i^b, \exists C_j \\ k_{N+1}, & \text{if } C_i \neq \text{ground AND } S_i = N + 1, \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

where k_{N+1} is an integer which corresponds to the cost of an invisible support.

The physical constraint energy E_{PC} is a weighted sum of C_C , C_{dist} and C_{SPC} because they have different influences in practice. The formal expression is:

$$E_{PC}(S, ST, C) = \alpha_C C_C(S_i, C_i) + \alpha_{dist} C_{dist}(S_i, C_i, ST_i) + \alpha_{SPC} C_{SPC}(S_i, C_i). \quad (12)$$

The optimal support relations are achieved by tuning the weights.

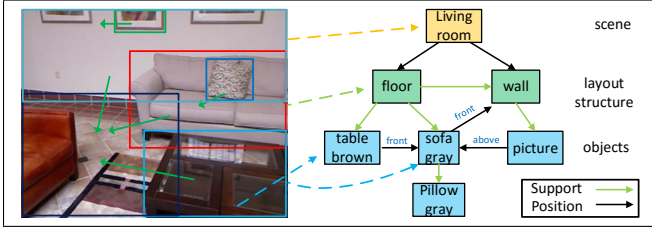


Fig. 4. An example of interpreting the image using our scene graph. In the root layer there is only one node to indicate the scene type; the first layer contains structure element ground, wall, ceiling and a hidden object; the lower layer contains other objects detected in the image except ground, wall, ceiling and hidden. Each node represents an individual object and each edge represents the support relation or relative position.

A. Energy minimization

The minimization of the energy function Eq. (5) can be formulated as an integer programming problem. Let $N^* = N + 1$ indicate the number of detected objects plus the hidden supports. The Boolean indicator variable $B_{SP_{i,j}}: 1 \leq i \leq N, 1 \leq j \leq 2N^* + 1$ encodes object C_i , its supporting objects C_j and $i \neq j$ and support type ST_i . $B_{SP_{i,j}} = 1, 1 \leq j \leq N^*$ means that object C_j supports object C_i from behind. If $N^* + 1 \leq j \leq 2N^*$, then object C_i is supported by C_{j-N^*} from below, and $j = 2N^* + 1$ indicates that C_i is the ground and need no support. Boolean variable $B_{C_{i,\lambda}} = 1$ indicates that object C_i has a class value λ . Furthermore, variable $\chi_{i,j}^{\lambda,v}$ encodes the case $B_{SP_{i,j}} = 1, B_{C_{i,\lambda}} = 1, B_{C_{j,v}} = 1$. The minimum energy inference problem is formulated as an integer program using this over-complete representation:

$$\arg \min_{B_{SP}, B_{C}, \chi} \sum_{i,j} \theta_{i,j}^{SP} B_{SP_{i,j}} + \sum_{i,\lambda} \theta_{i,\lambda}^C B_{C_{i,\lambda}} + \sum_{i,j,\lambda,v} \theta_{i,j,\lambda,v}^{\omega} \chi_{i,j}^{\lambda,v}. \quad (13)$$

In this formulation, the support energies E_{SP} in Eq. (7) and the distance constraints C_{dist} in Eq. (10) are encoded by $\theta_{i,j}^{SP}$; the object class energies E_C and support constraints C_{SPC} in Eq. (11) are encoded by $\theta_{i,\lambda}^C$; the object class constraints C_C in Eq. (9) are encoded by $\chi_{i,j}^{\lambda,v}$.

The support constraints C_{SPC} are enforced by

$$\sum_j B_{SP_{i,j}} = 1, \sum_j B_{C_{i,\lambda}} = 1, \forall i, \quad (14)$$

$$\sum_{j,\lambda,v} \chi_{i,j}^{\lambda,v} = 1, \forall i, \quad (15)$$

$$B_{SP_{i,j}} = B_{C_{i,\lambda}}, \text{ for } j = 2N^* + 1, \lambda = 1, \forall i. \quad (16)$$

To ensure the definition of $\chi_{i,j}^{\lambda,v}$ and satisfy the object class constraints C_C , we require that

$$\sum_{j,\lambda,v} \chi_{i,j}^{\lambda,v} = B_{SP_{i,j}}, \forall i, j \quad (17)$$

$$\sum_{j,\lambda,v} \chi_{i,j}^{\lambda,v} \leq B_{C_{i,\lambda}}, \forall i, \lambda. \quad (18)$$

The solution of the integer program is defined as:

$$B_{SP_{i,j}}, B_{C_{i,\lambda}}, \chi_{i,j}^{\lambda,v} \in \{0, 1\}, \forall i, j, \lambda, v. \quad (19)$$

Algorithm 1 Semantic Scene Graph Construction

```

1: Initialization:
2:  $root \leftarrow$  scene type
3:  $L \leftarrow root$ 
4: while there are unassigned objects do
5:   while  $L \neq 0$  do
6:      $parent \leftarrow$  first element of  $L$ 
7:     Remove first element of  $L$ 
8:     for each object supported by  $parent$  do
9:       Create node with
10:      Assign to parent
11:      Append  $L \leftarrow$  object
12:     end for
13:   end while
14:   Assign renaming objects to hidden node
15: end while
16: for  $i=1:N$  do
17:   for  $j=i:N$  do
18:     Connect  $v_i$  and  $v_j$  with edge  $e_{i,j}$ 
19:   end for
20: end for

```

To solve Eq. (19) is an NP hard problem. Therefore, we relax this equation as:

$$B_{SP_{i,j}}, B_{C_{i,\lambda}}, \chi_{i,j}^{\lambda,v} \in [0, 1], \forall i, j, \lambda, v. \quad (20)$$

Equation (20) is a linear program and can be solved by the LP solver of the Gurobi package.

V. SCENE GRAPH CONSTRUCTION

Given a set of detected objects $C = \{C_1, \dots, C_N\}$, the object positions $P = \{p_1, \dots, p_N\}$, attributes $A = \{A_1, \dots, A_N\}$, and relationships between objects $R = \{R_{i,j}, i \neq j\}$, a scene graph G is defined as a tuple $G = (V, E)$ where V is the set of vertices and E the set of edges, respectively. The triple $v_i \sim \{c_i, p_i, A_i\}$ represents object class c_i , position p_i in the scene and attributes A_i such as color, shape etc. We train classifiers for the 8 most familiar colors in our live: red, green, blue, yellow, brown, black, white and gray using RGB features to recognize the object color. The position information is $p_i = (b_i, z_{min}, z_{max})$, where b_i defines the bounding box of the object and (z_{min}, z_{max}) are its minimal and maximal depth respectively. The relationships $R_{i,j}$ represent support relations $T_{i,j}$ or relative positions between objects.

Fig. 4 shows an example of a scene graph construction for a given image in the left. It is constructed using support relations as explained in Algorithm 1. Please notice that a *hidden* vertex is added to the second layer, i.e. it is connected to the root node. Its purpose is that unsupported objects can be assigned to it. Furthermore, walls are supported by the ground by default, and the ceiling by the walls. For indoor scenes, these are reasonable assumptions.

In the next step, we need to connect objects so as to describe the relative position of two nodes. For each object, we only define spatial information for objects which are close instead of creating a fully connected graph. For example in

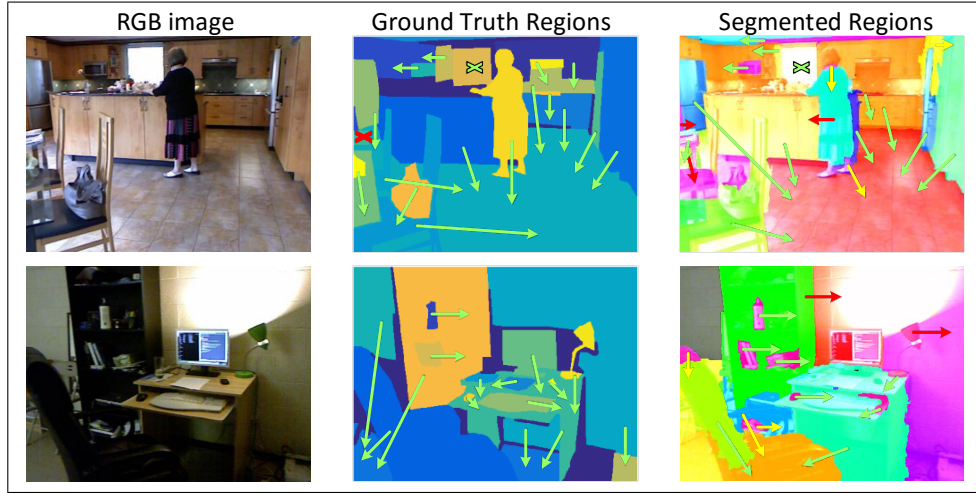


Fig. 5. Examples of support inference and object recognition. The middle column shows the results on the ground truth while the results in the last column are based on object detection. The direction of arrow indicates the supporting object. Cross denotes a hidden support. Correct support predictions in green and incorrect in red. The yellow ones mean that the support predictions are subjectively correct, but not exists in the ground truth. In the last column, objects belonging to the same class are denoted in the same color.

Fig. 4, the brown table can be described as in front of the sofa or in front of the wall. But the former one is more exact to describe this spacial relation and more useful for scene understanding than the latter. In this paper, we describe the relative position with concepts above/under, front/behind and right/left, and each pair of them are symmetric.

VI. EXPERIMENTS AND RESULTS

A. Dataset

Our experiments were conducted on the NYUv2 [9] dataset. This dataset consists of 1449 images in 27 indoor scenes, and each sample consists of an RGB image and a depth map. The original dataset contains 894 object classes with different names. But some of them are similar classes (e.g. table and desk), or present rarely in the whole dataset, which is difficult to manipulate in practice. Therefore, we merge the similar object classes and discard those that appear sporadically. Finally, the object set consists of 32 classes and 3 generalized classes as "other prop", "other furniture", and "other structure". The 9 most frequent scene types are selected and the rest are generalized into the 10th type of "others".

The dataset is partitioned into two disjoint training and testing subsets, using the same split as [9]. For evaluation of the generated scene graphs, we manually built a scene graph for each scene based on ground truth using our GUI.

B. Evaluating Support Relations

Object segmentation is the foundation of support relation inference. Therefore, we evaluate the proposed method on both the ground truth segmentation and our segmentation which is based on object recognition. In the case of an object being detected without any other objects next to it, this object is assigned to be supported by the nearest surface, and its supporting object class is counted as hidden. For some objects with complex shape and configuration, such

as a corner cabinet being hung on the walls, the support prediction is counted as correct whichever wall is its support. We also compare the experimental results with the best results of the most related work [9], [16].

The experimental results and the comparison are listed in Tab. I. The accuracy of support relations predictions differentiate between without and with support type. When the support type is not considered, the predicted support relations which have correct supporting and supported objects are counted as correct. When the support type is taken into account, only when the predicted support type (from behind or below) is also correct, this prediction is accounted to be correct. When using the ground truth, our method of using contextual relations between object categories outperforms using only 4 structure categories. It demonstrates the effectiveness of contextual relation between different classes of objects to understand their support relations. Comparing the results based on the ground truth and object recognition given by our approach, the latter performance drops about 23%. It explains that the accuracy of object recognition is the main limitation for understanding support relations between

TABLE I
RESULTS OF THE DIFFERENT APPROACHES TO PREDICT SUPPORT RELATIONS. THE ACCURACY IS MEASURED BY TOTAL SUPPORT RELATIONS WHICH IS CORRECTLY INFERRED DIVIDED BY THE NUMBER OF OBJECTS IN GROUND TRUTH. THE ABBR. SEG. IS SEGMENTATION.

Predicting Accuracy without Support Type			
Region Source	ground truth	initial seg.	object seg.
Silberman [9]	75.9	54.1	55.1
Xue [16]	77.4	56.2	58.6
Ours	88.4	\	65.7
Predicting Accuracy with Support Type			
Silberman [9]	72.6	53.5	54.9
Xue [16]	74.5	55.3	56.0
Ours	82.1	\	61.5

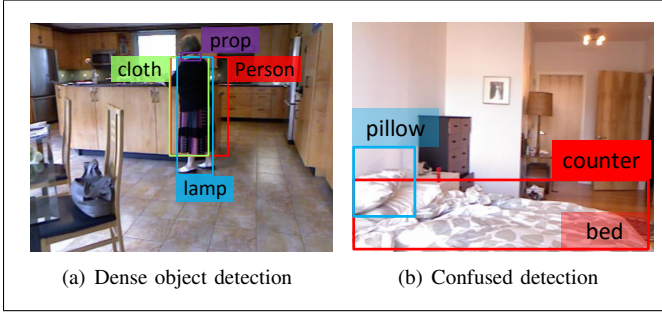


Fig. 6. Examples of imperfect object detection based on bounding box.

objects. This conclusion is verified again by the results given by [9] and [16] comparing the results based on the ground truth and their segmentation.

Our approach does not achieve significant boosting in support relation prediction using object segmentation against the other two works (the last column in Tab. I) for two reasons: 1) they used only 4 structure classes while our works detect 32 concrete object categories. We do not achieve very good results in object recognition; 2) the pixel-wise segmentation is more advantageous in estimating the spatial information about objects compared with our coarse segmentation. Furthermore, pixel-wise segmentation ensures that each object in given image has at least one support relation with one of its neighbor objects, while in our approach some objects are not detected, especially in a cluttered scene.

Nevertheless, only using the 2D bounding box and super-pixel maps is faster than previous works. In contrast, our method segments objects using the ready-made bounding boxes and superpixel maps provided by RCNN. No mutually call by each other of support relation and image segmentation [9], [10], which is the one of the main novelty in the other two works. Comparing the results of the two works between initial segmentation (the 3rd column) and improve object segmentation (the last column), their improved segmentation do not improve the predicting accuracy of support relations too much.

Visual examples are shown in Fig. 5. From the middle column we can see that, our approach performs well on the ground truth. The last column also illustrates the object recognition and segmentation. Some objects are not detected in the second row: the screen, keyboard and other props. Their regions are merged into other objects, because the bounding boxes involve them, e.g. the screen belongs to the wall region and the keyboard belongs to the table surface. The lamp is falsely considered to be supported by the wall because its joint lever is not detected. Another problem is that a complete object is sometimes recognized as multiple objects. For instance, the woman in the upper row is divided into 4 parts, cloth (cyan), body (blue), the neck is detected as a prop (pink) and the feet are detected as part of a lamp (light pink). It leads to incorrect support inference of the cloth being supported by the nearest cabinet. This phenomena is caused by dense detection in the regions of the person, as shown in Fig. 6(a).

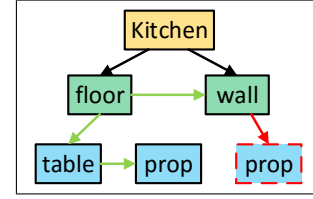


Fig. 7. An example of scene graph with an error (in red).

TABLE II
MATRIX OF THE SCENE GRAPH. THE RED NUMBERS INDICATE THE ERRORS FROM THE ABOVE GRAPH.

		Supporting				
		kitchen	floor	wall	table	prop
Supported	kitchen	0	1	1	0	0
	floor	1	0	0	0	0
	wall	1	0	0	0	0
	table	0	1	0	0	0
	prop	0	0	0/1	1/0	0

Because our inference is jointly minimized by the energy function Eq. (6), the final results of the object recognition accuracy are improved. For instance as shown in Fig. 6(b), the white bed with large flat surface is recognized as counter and bed in the same bounding box. Due to a pillow supported by this confused object, it leads the inference to decide it as bed, because a pillow is rarely supported by a counter in the prior knowledge.

C. Evaluating Scene Graph

Because our goal is to evaluate the structure quality of generated scene graph, the attributes of objects and relative positions between objects are not taken into account in this work. We represent the directed, unweighted graph by its affinity matrix, as shown in Fig. 7 and Tab. II. Here, the object class corresponding to the columns supports the classes corresponding to the rows.

To measure the similarity with ground truth graphs, we create graphs $G'(V', E')$ with undirected edges such that $V' = V$ and $E' = \{e'_{ij} = 1 \Leftrightarrow e_{ij} = 1 \vee e_{ji} = 1\}$. We do so for both the estimated scene graph and the ground truth. To estimate their similarity, we compare their Cheeger constants. The Cheeger constant h_G of a graph is defined to be $h_G = \min_S h_g(S)$ [25]. Here, S denotes a subset of the vertices of G , and

$$h_G(S) = \frac{|E(S, \bar{S})|}{\min(\text{vol}S, \text{vol}\bar{S})} \quad (21)$$

with $\text{vol}S = \sum_{x \in S} d_x$ being the volume of S . d_x is the degree of vertex x , and \bar{S} is the complement set of S , i.e. $\bar{S} = V \setminus S$. The symbol $|\cdot|$ indicates the cardinality. Since h_G is hard to compute, we use upper and lower bounds

$$l_G = \frac{1}{2}(1 - \lambda_2) \leq h_G \leq \sqrt{2 - 2\lambda_2} = u_G. \quad (22)$$

Here, λ_2 denotes the second largest eigenvalue of the random walk matrix $P(G) = D^{-1}A$ of the graph G with affinity matrix

TABLE III

RESULTS OF DIFFERENT MEASURES TO EVALUATE THE QUALITY OF GENERATED SCENE GRAPH COMPARING WITH GROUND TRUTH. THE NUMBER IS SMALLER, THE QUALITY IS BETTER. 0 MEANS THE GENERATED GRAPH IS IDENTICAL WITH THE GROUND TRUTH.

Evaluation of generated scene graph			
Measures	Cheeger (22)	Spectral (23)	Naive
Mean	0.19	0.20	0.41
Variance	0.16	0.07	0.15

A and $D_{ii} = d_x$. We then take $|(u'_G - l'_G) - (u_{H'} - l_{H'})|$ as similarity between G' and the undirected graph H' corresponding to the ground truth graph H .

Since there may be incorrectly estimated scene graphs whose Cheeger constants nonetheless do not differ from those of the ground truth graph², we take

$$\left\| u_2(G') \cdot u_2(G')^\top - u_2(H') \cdot u_2(H')^\top \right\|_F / \sqrt{|V(H')|} \quad (23)$$

as further measure of the similarities between the two graphs. In Eq. (23), u_2 denotes the eigenvector corresponding to λ_2 and $\|\cdot\|_F$ the Frobenius-norm, $|\cdot|$ the cardinality, and $\sqrt{|V(H')|}$ is for normalization.

Lastly, we use a naive heuristic to measure the difference between the two graphs. Its computation is explained in the supplementary material.

The evaluation results using the three different measures on the test dataset are shown in Tab. III. We can see that the generated scene graphs achieve low mean error values when comparing with ground truth. It proves that our method generates reasonable scene graph given a scene.

VII. CONCLUSION

This work presents a new approach for inferring accurate support relations between objects from given RGBD images of cluttered indoor scenes. We also introduce how to construct semantic scene graphs interpreting physical and contextual relations between objects and environment. This topic is a necessary step for deeper scene understanding. The proposed framework takes RGBD images as input, detects object using RCNN and then conducts a simple object extraction from a superpixel map, which is faster than pixelwise segmentation. Next, reasonable support relations are inferred by using physical constraints and prior support knowledge between object classes. Finally, support relations, along with contextual semantics, scene recognition, and object recognition allow to infer a semantic scene graph. Using the NYUv2 dataset, the inferred support relationships are more accurate than those achieved from previous algorithms. For assessing the semantic scene graphs, ground truth graphs are created, and objective measures for graph comparison are proposed. Evaluation results show that the inferred scene graphs are reasonable. The ground truth graphs and the tool to create them will be public available.

In future research, we will experiment on letting robot finish specific tasks using our scene graphs. Furthermore, it

would be nice to give the support relations between objects semantic meaning, e.g. "the man is sitting in the sofa". From the scene graph, it would be interesting to group objects into meaningful groups, such as studying area. The scene graphs also should improve the prediction accuracy of support relations and object classes in an iterative manner. At last, we will improve the measures to assess the quality of the scene graph hypotheses.

REFERENCES

- [1] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.
- [2] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese, "Understanding indoor scenes using 3d geometric phrases," in *CVPR*, 2013, pp. 33–40.
- [3] R. Mottaghi, M. Rastegari, A. Gupta, and A. Farhadi, "What happens if... learning to predict the effect of forces in images," *arXiv preprint arXiv:1603.05600*, 2016.
- [4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [5] C. Wu, I. Lenz, and A. Saxena, "Hierarchical semantic labeling for task-relevant rgb-d perception," in *Robotics: Science and systems*, 2014.
- [6] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena, "Semantic labeling of 3d point clouds for indoor scenes," in *Advances in Neural Information Processing Systems*, 2011, pp. 244–252.
- [7] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *CVPR*, 2015, pp. 3668–3678.
- [8] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proceedings of the Fourth Workshop on Vision and Language*, 2015, pp. 70–80.
- [9] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*. Springer, 2012, pp. 746–760.
- [10] F. Xue, S. Xu, C. He, M. Wang, and R. Hong, "Towards efficient support relation extraction from rgbd images," *Information Sciences*, vol. 320, pp. 320–332, 2015.
- [11] N. Prabhu and R. Venkatesh Babu, "Attribute-graph: A graph based approach to image ranking," in *ICCV*, 2015, pp. 1071–1079.
- [12] D. Lin, S. Fidler, C. Kong, and R. Urtasun, "Visual semantic search: Retrieving videos via complex textual queries," in *CVPR*, 2014, pp. 2657–2664.
- [13] T. Liu, S. Chaudhuri, V. G. Kim, Q. Huang, N. J. Mitra, and T. Funkhouser, "Creating consistent scene graphs using a probabilistic grammar," *ACM Transactions on Graphics*, vol. 33, no. 6, p. 211, 2014.
- [14] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *arXiv preprint arXiv:1602.07332*, 2016.
- [15] Z. Jia, A. Gallagher, A. Saxena, and T. Chen, "3d-based reasoning with blocks, support, and stability," in *CVPR*, 2013, pp. 1–8.
- [16] B. Zheng, Y. Zhao, J. Yu, K. Ikeuchi, and S.-C. Zhu, "Scene understanding by reasoning stability and safety," *IJCV*, vol. 112, no. 2, pp. 221–238, 2015.
- [17] Y.-S. Wong, H.-K. Chu, and N. J. Mitra, "Smartannotator an interactive tool for annotating indoor rgbd images," in *Computer Graphics Forum*, vol. 34, no. 2. Wiley Online Library, 2015, pp. 447–457.
- [18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
- [19] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*. Springer, 2014, pp. 345–360.
- [20] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics*, vol. 23, no. 3. ACM, 2004, pp. 309–314.
- [21] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *ICCV*, 2009, pp. 277–284.

²Please refer to the supplementary material for an example.

- [22] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015, pp. 3431–3440.
- [23] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *CVPR*, vol. 2, 2006, pp. 2169–2178.
- [24] J. M. Coughlan and A. L. Yuille, “Manhattan world: Orientation and outlier detection by bayesian inference,” *Neural Computation*, vol. 15, no. 5, pp. 1063–1088, 2003.
- [25] F. Chung, *Spectral graph theory (CBMS regional conference series in mathematics, No. 92)*. American Mathematical Society, 1996.

VIII. SUPPLEMENTARY MATERIAL

A. Support Features

Support features are listed in Table IV. Some examples of inferring support relations in given images are shown in Fig. 8.

B. Relative Position Decision

For object i , its spatial information can be represented by 3D room coordinates $(x_{min}^i, x_{max}^i, y_{min}^i, y_{max}^i, z_{min}^i, z_{max}^i)$, where $x-z$ is the floor plane and y direction points upward. For convenient discussion, we define some symbols here. $I_{i,j}^x = [x_{min}^i, x_{max}^i] \cap [x_{min}^j, x_{max}^j]$ denotes the intersection between object i and j when they are projected on the x axis, and so do $I_{i,j}^y$ and $I_{i,j}^z$ too. The rule for deciding object i position relative to object j is listed in Table V and illustrated in Fig. 9.

Because the position of above-under, behind-front and right-left are symmetric, we don’t describe the versa here any more.

C. Evaluating Scene Graph

1) *Laplacian Measure*: Consider the examples of scene graphs shown in Fig. 11. In the left plot, a ground truth scene graph is shown. The right image in the same figure shows an example of an estimated graph. It can be easily seen that the subgraph with root being the *table* vertex is incorrectly assigned to the *wall*. In the following, consider the graphs relaxed so as to have undirected edges.

A graph measures based on an isoperimetry such as Eq. (18) cannot capture this difference since the ratio between surface area and volume remains unchanged.

Therefore we further use a measure inspired by a normalized cut of each graph given by the eigenvector u_2 to the second smallest eigenvalue of the graph Laplacian. For the left graph the decision boundary induced by the hyperplane with normal u_2 cuts the graph between *kitchen* and *floor*, whereas it cuts between *kitchen* and *wall* for the graph shown in the right image of Fig. 11. A measure of the difference between the two hyperplanes is given by

$$\|P_{u_2(G_1)} - P_{u_2(G_2)}\|_F \quad (24)$$

Relative Position	Condition
Above	$y_i^{min} > y_j^{max}; z_i^{min} < z_j^{min} < z_j^{max}; I_{i,j}^x, I_{i,j}^z \neq \emptyset$
Behind 1	$z_i^{min} \geq z_j^{max}; I_{i,j}^x \neq \emptyset$
Behind 2	$\frac{1}{2}(z_i^{min} + z_i^{max}) > z_j^{max}; I_{i,j}^x \neq \emptyset$
Right	$z_{min}^j < \frac{1}{2}(z_i^{min} + z_i^{max}) < z_{max}^j; \frac{1}{2}(x_i^{min} + x_i^{max}) > x_{max}^j$

TABLE V

THE DECISION RULES OF OBJECT POSITION RELATIVE TO AN OBJECT.

Supported	Supporting							
	kitchen	floor	wall	table	chair	picture	cup	book
	kitchen	0	1	1	0	0	0	0
	floor	1	0	0	0	0	0	0
	wall	1	0	0	0	0	0	0
	table	0	1	0	0	0	0	0
	chair	0	1	0	0	0	0	0
	picture	0	0	1	0	0	0	0
	cup	0	0	0	1	0	0	0
	book	0	0	0	1	0	0	0

TABLE VI

MATRIX OF THE GROUND TRUTH SCENE GRAPH FIG. 11(A).

		Supporting							
Supported		kitchen	floor	wall	table	chair	picture	cup	book
	kitchen	0	1	1	0	0	0	0	0
	floor	1	0	0	0	0	0	0	0
	wall	1	0	0	0	0	0	0	0
	table	0	0	0	0	0	0	0	0
	chair	0	1	0	0	0	0	0	0
	picture	0	0	1	0	0	0	0	0
	cup	0	0	0	1	0	0	0	0
	book	0	0	0	1	0	0	0	0

TABLE VII

MATRIX OF THE CONSTRUCTED SCENE GRAPH FIG. 11(B). THE RED NUMBERS INDICATE THE ERRORS IN THE GRAPH: THE TABLE IS SUPPORTED BY WALL INSTEAD OF BY FLOOR.

where $u_2(G_i)$ denotes the second smallest eigenvector of the Laplacian of graph G_i , and P_{u_2} the orthogonal projection onto $\text{span}(u_2(G_i))$. Since the number of vertices can differ between images, we normalize Eq. (24) by the maximum number of vertices the ground truth scene graphs can have.

2) *Heuristic*: Beside the two evaluation methods “Cheeger” Eq. (19) and “Spectral” Eq. (20) as described in the paper, we propose another naive method to measure the constructed graph. The matrices for describing the ground truth scene graph Fig. 11(a) and the constructed graph Fig. 11(b) are shown in Tab. VI and Tab. VII respectively. The matrix can not only describe the support relations between object but also the object classification. The difference between matrices M_i and M_j is formally calculated as:

$$d_{i,j} = \frac{|M_i \oplus M_j|}{|M_i \vee M_j|} \quad (25)$$

where $|\cdot|$ is the total number of 1 in a matrix. Even though it is not a sophisticated method, it is a complementary measure for reference.

D. GUI

We provide a convenient GUI tool for generating ground truth graphs. Upon acceptance, the tool and the ground truth scene graph for NYUv2 dataset will be available on the authors’ home page. Please turn to the video supplementary material to have a look at our GUI tool.

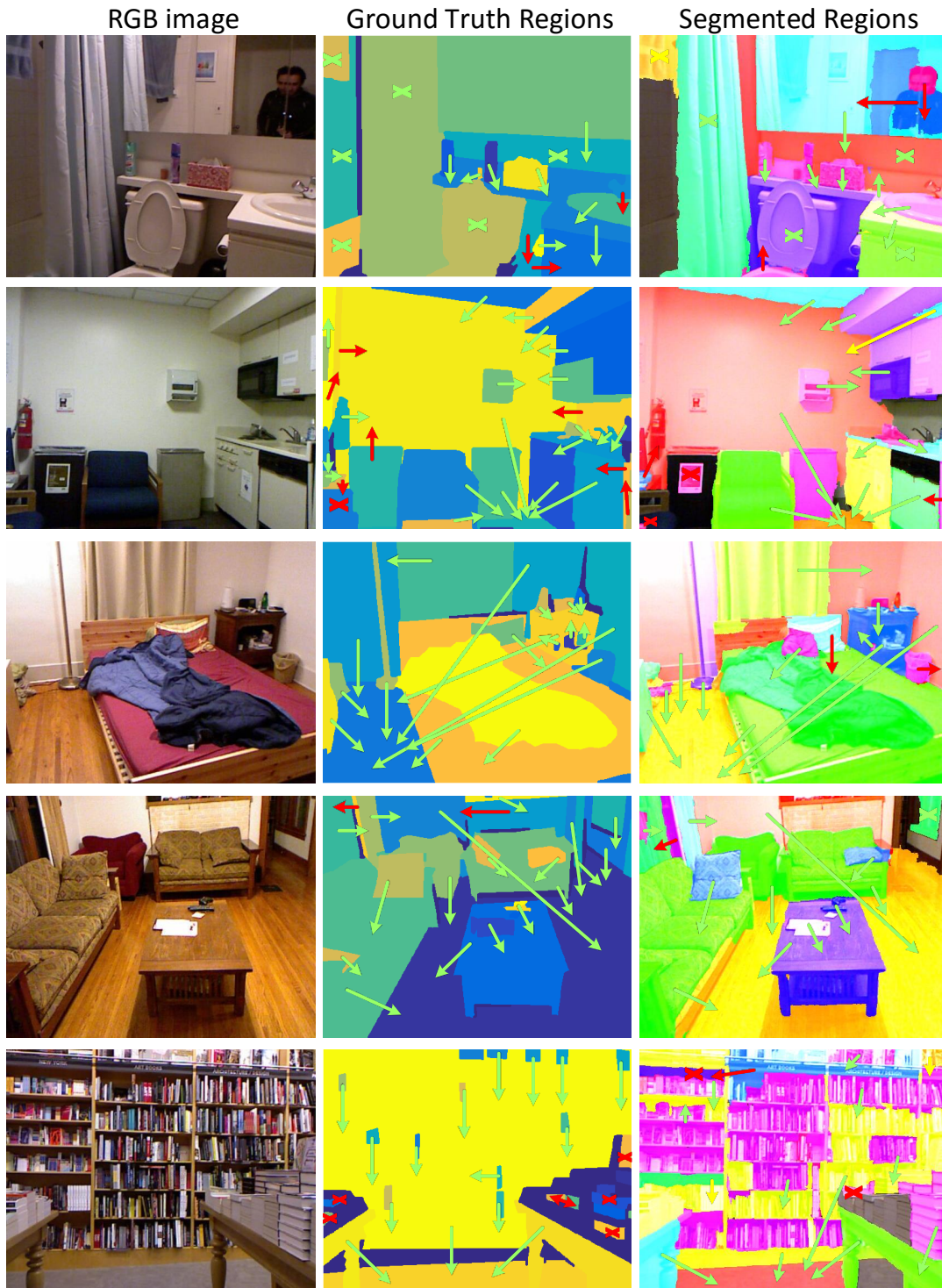


Fig. 8. Examples of support and object class inference with the LP solution. The middle column shows the results on the ground truth while the results in the last column are based on object detection. The direction of arrow indicates the supporting object. Cross denotes a hidden support. Correct support predictions in green and incorrect in red. The yellow ones mean that the support predictions are subjectively correct, but the object detection or segmentation are incorrect. In the last column, objects belonging to the same class are denoted in the same color.

Support Feature Description	Dims
Geometry	8
G1. Minimum vertical and horizontal distance between the two regions	2
G2. Absolute distance between the regions' centroids	1
G3. The lowest heights of the two regions above the ground	2
G4. Percentage of the supporting region that is farther from the viewer than the supported region	1
G5. Percentage of the supported region contained inside convex hull of supporting region's projection onto the floor plane	1
G6. Percentage of the supported region contained inside convex hull of supporting region's horizontal points when projected onto the floor plane	1
Shape	9
S1. Number and percentage of horizontal pixels in the supporting region	2
S2. Number and percentage of horizontal pixels in the supported region	2
S3. Number and percentage of vertical pixels in the supporting region	2
S4. Number and percentage of vertical pixels in the supported region	2
S5. Chi-squared points when projected onto the floor plane	1
Region	3
R1. Ratio of number of pixels between the supported supported and supporting region	1
R2. Whether or not the two regions are neighbors in the image plane	1
R3. Whether or not the supporting region is hidden	1

TABLE IV
SUPPORT FEATURE DESCRIPTION.

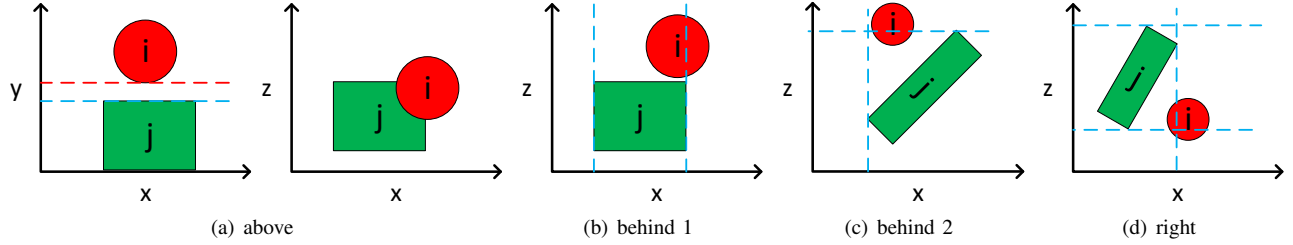


Fig. 9. Relative position described by (a) above; two case of behind in (b) and (c); and right in (d).

